

**TASK**

**Exploratory Data Analysis on the Athletes Data Set**

[](https://www.hyperiondev.com/)

**SUMMARY OF THE DATA SET**

Data set is made of 301 rows and 8 columns labelled S.NO, Name, Nationality, Current Rank, Previous Year Rank, Sport, Year, earnings ($ million).

Entries are grouped by year then ranked within that year group.

**DATA CLEANING**

0 duplicate rows are found.

Variations of the same sport are entered. Also, some categories are made separate when in fact they are a subcategory of the general sport.

There is one row with the name Deion Sanders and it has an inconsistent value for Sport column i.e American football/baseball. After researching by the name to find out the actual sport played, the column value was changed to soccer.

Overall unique values are reduced to basketball, soccer and American football- which are 2 different sports, tennis, hockey, golf, auto racing, boxing.

Auto racing can be both Nascar and F1 e.g Lewis Hamilton is in F1 but the value entered for the sport he plays is auto racing. As there are too many entries to search for manually to change the column value to either Nascar or F1, Nascar and F1 racing included in one category i.e auto racing. Similarly, information about motorcycle gp and motorsports is included in auto racing.

There are entries of basketball which are also NBA, e.g. Kobe Bryant and there is one entry of NBA alone by Shaquille O'Neal. Hence, NBA is added to basketball.

Entries with NFL are not found in American football, but American football contains sportsmen who played for NFL e.g. Eli Manning. NFL is included in American football.

**MISSING DATA**

24 rows out of a total of 301 have missing values for previous year rank. Rows with NaN are removed before included later in analysis.

**DATA STORIES AND VISUALISATIONS**

The pie chart shows the proportion of sportsmen per country. USA has the highest share while Mexico has the lowest.

Chart, pie chart

Description automatically generated

The number of sportsmen by sport is illustrated below. Most sportsmen listed in Forbes play basketball, however basketball is not the highest paying on average, as the second figure above shows.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

The boxplot below shows that boxing has the highest outliers for earnings. Excluding outliers, soccer has the highest maximum value and also highest upper quartile of earnings out of all sports except MMA. The minimum value of earnings for boxing, golf, basketball shows they are on par. Soccer also has the second highest median earnings and MMA has a higher single value which is also a median.

Chart, box and whisker chart

Description automatically generated

Tiger Woods and Michael Jordan were listed the most.

Graphical user interface, text, application, email

Description automatically generated

Total earnings for all sports over time increased overall from 1990-2020, assuming the real value of money was inputted for each year.

Chart, line chart

Description automatically generated

Similarly, Cristiano Ronaldo’s total earnings over time increased.

Chart, line chart

Description automatically generated

Top 10 highest paid sportsmen. Trend of small increments in earnings from the previous sportsman excl. Tiger Woods who has higher increment in income than the average for the rest of the sportsmen.

Chart

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Higher earnings for US nationals due to longer tail of distribution (red graph) but very low density is represented. There is higher density between 55-175 for non-US so more get higher earnings. Highest density is displayed for US, top density for US shows lower earnings than top density for non-US.

Chart

Description automatically generated

**THIS REPORT WAS WRITTEN BY: ROBERTA DOBRA**

**STUDENT NUMBER: RA23010006214**

